The Risk of Disclosure When Reporting Commonly Used Univariate Statistics

Ben Derrick(✉), Elizabeth Green, Felix Ritchie, and Paul White

University of the West of England, Bristol, UK
ben.derrick@uwe.ac.uk

Abstract. When basic or descriptive summary statistics are reported, it may be possible that the entire sample of observations is inadvertently disclosed, or that members within a sample will be able to work out responses of others. Three sets of univariate summary statistics that are frequently reported are considered: the mean and standard deviation; the median and lower and upper quartiles; the median and minimum and maximum. The methodology assesses how often the full sample of results can be reverse engineered given the summary statistics. The R package `uwedragon` is recommended for users to assess this risk for a given data set, prior to reporting the mean and standard deviation. It is shown that the disclosure risk is particularly high for small sample sizes on a highly discrete scale. This risk is reduced when alternatives to the mean and standard deviation are reported. An example is given to invoke discussion on appropriate reporting of summary statistics, also giving attention to the box and whiskers plot which is frequently used to visualise some of the summary statistics. Six variations of the box and whiskers plot are discussed, to illustrate disclosure issues that may arise. It is concluded that the safest summary statistics to report is a three-number summary of median, and lower and upper quartiles, which can be graphically displayed by the literal 'boxplot' with no whiskers.

Keywords: SDC · Statistic · Disclosure · Control · Summary · Quartile · Boxplot

1 Introduction

In statistical analyses there is potential conflict between providing useful results and protecting the confidentiality of individuals within the data [1]. Given commonly reported univariate summary statistics, it may be possible to construct the exact frequencies of values within a sample, which in many contexts would be unwarranted disclosure.

For illustrative purposes, consider a four-point scale for reporting health on a survey (1 = Good health, 2 = Fair health, 3 = Bad health, 4 = Very bad health). In a summary of

J. Domingo-Ferrer and M. Laurent (Eds.): PSD 2022, LNCS 13463, pp. 119–129, 2022.
https://doi.org/10.1007/978-3-031-13945-1_9

the results separated by gender and ethnicity, assume the following means and standard deviations (SD) for males are reported:

White: N = 18 Mean = 2.06 SD = 0.998
Mixed: N = 8 Mean = 2.00 SD = 0.926
Asian: N = 6 Mean = 2.67 SD = 0.816
Black: N = 5 Mean = 2.00 SD = 0.000
Other: N = 1 Mean = 5.00 SD = 0.000

There is debate regarding ascribing numeric values to ordinal data for analyses. This is frequently done in practice, and is might not be unreasonable when pragmatic assumptions of equal distance between groups are stated [2].

In the example, the ethnic groups with standard deviation equal to zero must all have reported the same value (the group mean). It may not be as straightforward to reverse engineer the frequencies for the remaining three groups, but the R package `uwedragon` will show the plausible frequency distributions for a given sample size, mean and standard deviation [3]. For the 'Asian' and 'Mixed' groups there are only two frequency distributions possible with the stated means and standard deviations. For the 'White' group there are four frequency distributions possible. Table 1 is the table of results as given by Lowthian and Ritchie [4].

Table 1. Health survey responses for males by ethnicity.

	Good	Fair	Bad	Very bad	Total
White	6	7	3	2	18
Mixed	2	2	3	1	8
Asian	1	0	5	0	6
Black	0	5	0	0	5
Other	0	0	0	1	1
Total	9	14	11	4	38

As Lowthian and Ritchie [4] state, from this table we draw several conclusions:

- The single male who does not identify with any of the ethnic groups has 'Very bad health'. This can cause group attribute disclosure but not necessarily reidentification.
- All of the individuals who identify as Black have 'Fair health'.
- The one Asian who responded that he enjoys 'Good health' knows that his Asian colleagues all report 'Bad health'.

The reporting of mean and standard deviation in addition to, or as an alternative to, the frequency table may also have similar associated disclosure risk. As stated, there are two possible solutions for the Asian category with the given mean and standard deviation. The R package `uwedragon` shows that sample values are either {1, 3, 3, 3, 3, 3} or {2, 2, 2, 3, 3, 4}. Thus, again if an individual within this grouping reported 1 'Good health', he can work out that his Asian colleagues all reported 3 'Bad health'. If in fact the second solution had been true, then the person reporting a 4 'Very bad health' would know that all of the other respondents in that grouping had reported better health.

The `uwedragon` package can help identify the level of risk by supplying detail of the possible solutions for a given sample size, mean and standard deviation. Furthermore, the `uwedragon` package offers suggestions for disguising the mean and standard deviation when the risk level is high but there is still a need to report these figures [3]. The addition of noise in this way, or a similar manner, reduces the risk of reconstruction [5].

It may be that there are other less disclosive summary statistics that could be alternatively reported. An alternative measure of location to the mean could be the median. Likewise, an alternative measure of variability to standard deviation would be interquartile range, taken from lower quartile and upper quartile. These alternatives are based simply on an ordered location point, so will result in a reduced capacity to reconstruct an entire set of values or identify extreme observations. Algorithms for estimating the mean and standard deviation based on actual median, range and sample size can be utilized [6]. Reporting median, range and sample size alongside estimates for the mean and standard deviation would have a disguise effect reducing the risk of reconstruction.

This paper provides methodology and results that raise awareness of the potential disclosure risk when reporting only the mean and standard deviation, particularly for small measurement scales and small sample sizes. This paper further considers the use of alternative summary statistics that may be less disclosive in these situations. The alternative summary statistics considered are: either only the median, lower quartile and upper quartile; or only the median, minimum and maximum.

2 Methodology

We consider a scale restricted to k defined points for a sample of size n. The total sample space is the number of combinations for the values 1 to k in a sample of size n. Univariate summary statistics of each combination within the sample space are calculated and compared to the same summary statistics for each other combination within the sample space. A high proportion of combinations within the sample space that can be uniquely identified by the given summary statistics is a high disclosure risk.

For example, for a $k = 5$-point scale with $n = 5$, the total sample space is 129. The combination of sample values $\{1, 2, 2, 3, 4\}$, has mean $\bar{x} = 2.40$ with standard deviation $s = 1.14$. No other combination within the sample space gives this same $\bar{x}$ and s, and this is referred to as a unique identification. In fact, 87 of the 129 possible combinations for k = 5 and n = 5 can be uniquely identified by their mean and standard deviation.

Using the approach by Derrick et. al. [7], we report the total number of possible different sample configurations for sample sizes $n = 3, 4, 5, \dots 10, 11, 12$. We then report the number of these samples which can be uniquely identified through knowing the mean and standard deviation when reported with full precision, and when reported to two decimal places or one decimal place (divisor of variance used = $n - 1$).

A summary of the results is given for a 7-point scale and a 10-point scale (with additional scales in the appendix). These tables summarise those situations where there is a unique one-to-one correspondence between $(\bar{x}, s, n)$ and a sample configuration leading to $(\bar{x}, s, n)$ uniquely identifying the sample which gives rise to $(\bar{x}, s)$.

This methodology is herein extended to give the number of unique solutions when alternative summary statistics are given. Firstly, if only the median, first quartile (Q1) and third quartile (Q3) are reported. Secondly, if only the median, minimum and maximum are reported.

The minimum and maximum values are the true minimum and true maximum from the sample. The median is calculated as the middle value in the ordered sample (or mid-point of two central values if sample size is an even number). The calculation of quartiles differs in common statistical software. We consider several of these approaches calculated using the `quantile` function in R [8]. Mathematical definition of the methods is given by Hyndman and Fan [9]. SPSS and Minitab both use 'method 6', the R default is 'method 7', whereas the SAS default is 'method 2'.

3 Results

Table 2 gives the number of unique identifications for the given summary statistics reported when data is from an inherent 7-point scale. Table 3 provides the same information for a 10-point scale.

Table 2. Number of unique solutions, data on 7-point scale.

n	Sample space	Mean and SD reported			Median, Q1 and Q3 reported			Median, Min, Max
		Full	2dp	1dp	SPSS	R	SAS	reported
3	84	76	76	76	84	84	84	84
4	210	143	143	143	210	210	180	85
5	462	206	193	193	80	7	7	7
6	924	246	222	200	440	28	4	13
7	1716	295	253	203	0	24	0	7
8	3003	289	289	201	59	59	16	13
9	5005	405	325	215	0	0	0	7
10	8008	438	361	202	3	59	0	13
11	12376	493	397	198	0	0	0	7
12	18564	533	433	213	3	3	0	13

Reporting the median, Q1 and Q3, theoretically contains a smaller number of disclosive scenarios than reporting the mean and SD, when $n > 6$. By virtue of reporting quantiles to decimals of 0 or 0.5 as per method 2 in [9], the approach to calculating quantiles adopted by SAS is the least disclosive, relative to procedures in Minitab, SPSS and R.

Table 3. Number of unique solutions, data on 10-point scale.

n	Sample space	Mean and SD reported			Median, Q1 and Q3 reported			Median, Min, Max
		Full	2dp	1dp	SPSS	R	SAS	reported
3	220	188	188	188	220	220	220	220
4	715	353	353	343	705	705	468	181
5	2,002	509	422	346	128	10	10	10
6	5,005	564	472	332	1072	52	4	19
7	11,440	747	527	310	0	36	0	10
8	24,310	603	603	344	64	64	16	19
9	48,620	955	676	310	0	0	0	10
10	92,378	944	749	338	0	64	0	19
11	167,960	1134	822	286	0	0	0	10
12	293,930	1143	895	291	0	0	0	19

Reporting the maximum, minimum and median has low disclosive risk in terms of the entire set of sample values being revealed, particularly if the maximum or minimum value within a sample is not unique. However, there may be serious misgivings in practice regarding reporting the minimum and maximum. Paradoxically, revealing these values may protect the rest of the sample from being revealed.

The unique solutions for n > 4 when reporting the median, minimum and maximum represent the cases where all sample values are identical. Due to the standard deviation of zero, such combinations are also identifiable if the mean and standard deviation are instead reported. However, reporting the median, 1st quartile and 3rd quartile in these instances does not necessarily reveal all sample values.

When reporting the mean and SD, if the sample space is large, i.e. $k \geq 10$ and $n \geq 10$, the percentage of times the true underpinning sample is discovered is less than 1%.

Summary statistics assessed above may be reported in different combinations. The methodology could be extended to numerous different statistical reporting combinations. For example, the default descriptive statistics option in SPSS leads users to report all of the univariate summary statistics above and also include statistics for skewness and kurtosis. Note that adding additional summary statistics will increase the disclosure risk. For instance, reporting mean and standard deviation with the median, will result in a higher number of unique combinations being revealed than reporting only the mean and standard deviation.

4 Discussion Example

Consider the following hypothetical set of exam marks {1, 40, 50, 55, 58, 58, 60, 62, 65, 66, 66, 68, 70, 71, 72, 74, 75, 75, 80, 85}. Here there is a duty not to reveal individual exam marks when summarising the results, and particular care will have to be taken regarding the lowest scorer.

In this scenario the mean = 62.55 and SD = 17.90. You may take some comfort that the sample size and possible scale combination is too large for the `uwedragon` package to identify possible distributions of results. However, the high value of the standard deviation indicates the presence of some extreme values, and work could commence on identifying possible maximum and minimum values [7].

The median is midway between the 10^{th} and 11^{th} observation = 66. How the quartiles differ depending on approach used is shown in Table 4. The different approaches to calculating the quartiles could offer further assistance to protecting the data from being reversed engineered to reveal all values, if the method is chosen at random and not reported to the end user. Reporting Q1 and Q3 gives an idea of the spread of the data without revealing information about any potential extreme observations.

Table 4. Calculation of quartiles.

R function	1st quartile	3rd quartile	Note
Quantile (type = 2)	58	73.0	As per SAS
Quantile (type = 6)	58	73.5	As per Minitab & SPSS
Quantile (type = 7)	58	72.5	Default in R
Summary	58	72.5	
Fivenum	58	73.0	
Boxplot	58	73.0	With true min/max

The true minimum is 1, and the true maximum is 85. Note that some statistical software may present alternative minimum and maximum values with subsequent reporting of 'mild' or 'extreme' outliers. However, values in the extremes may be sensitive information, which may not be appropriate to disclose. Here, reporting the minimum is a risk of revealing that the weakest performer scored 1/100 on the exam. Likewise, for scales where there is no upper limit (e.g. salary) it may be more appropriate to report the upper quartile rather than the maximum.

5 Graphical Representation

Graphically, the summary statistics considered in examples like the above are often displayed in a box and whiskers plot, so natural temptation may be to summarise the data in this way. These depictions are often described as a five-number summary of minimum, lower quartile, median, upper quartile and maximum. However, this description is not always entirely accurate, and in fact can disclose many more than five values when 'outliers' are present.

Six variations of box-and-whisker plots or 'boxplots' are considered. Illustrations of each of the variations are given in Figs. 1 through Fig. 6, for the discussion example data. Included below is a description of each variation with a statement of causes for concern. The graphics, including the applicable quartile calculation, are as per the `boxplot` function in R [8].

1 **Tukey's schematic plot.** This is the traditional box and whiskers plot with interquartile range (IQR) $\times$ 1.5 for whiskers [10].

- Extreme observations are explicitly revealed.
- For each 'outlier' that is revealed, in addition to values for the 'minimum' and 'maximum', it slightly increases the opportunity for the sample to be reconstructed by a determined individual.

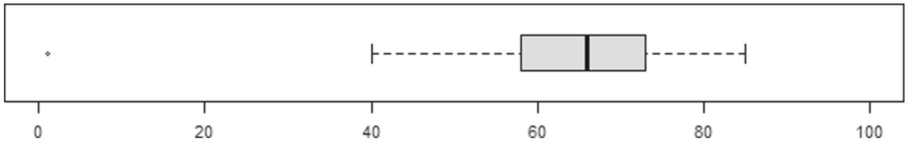

Fig. 1. Turkey schematic plot, traditional box and whiskers plot with IQR*1.5 for whiskers

2 **Box and whiskers plot with mean inserted**

- Reporting both the mean and median may give an indication of the direction and magnitude of extreme observation/s, even if outliers are removed from the plot the position of the mean relative to the median alludes to these extremes.
- The reporting of additional summary statistics increases the opportunity for the sample to be reconstructed by a determined individual.

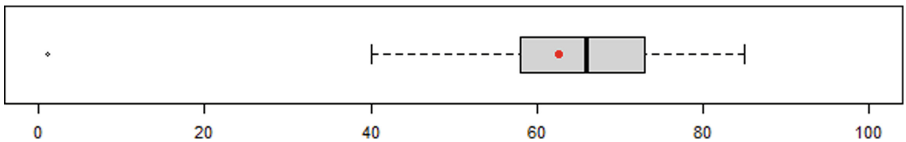

Fig. 2. Box and whiskers plot with mean added

3 **Modifications to the traditional calculation of the whiskers within a box and whiskers plot**

- Same issues as the traditional box and whiskers plot, but even more observations are explicitly revealed if the multiplier for IQR is <1.5.

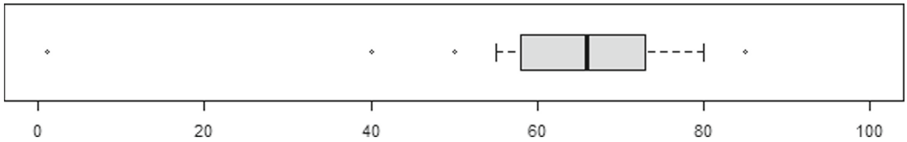

Fig. 3. Modification to traditional calculation IQR*0.5 for whiskers

4 **Box and whiskers plot using true minimum and maximum**. In this scenario whiskers are not calculated based on IQR, but extend to the full range of the data.

- Maximum and minimum explicitly revealed.
- Distorted impression of distribution if maximum or minimum is an extreme outlier.

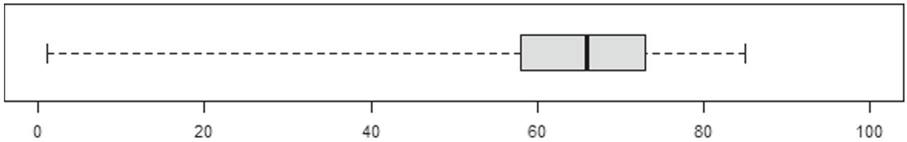

Fig. 4. Box and whiskers plot unmodified for extreme values

5 **Unstapled box and whiskers plot.** Here whiskers are calculated as per Tukey [10], but outliers are removed. The staples are subsequently removed herein to indicate that there may be extreme values beyond the reach of the whiskers.

- The most extreme observations are not explicitly revealed, but individuals within these missing extremes will be aware that they are 'outliers'.
- Without clear statement of the form of boxplot, incorrect perception of the true maximum and true minimum is possible.

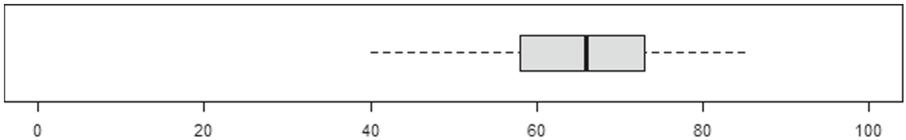

Fig. 5. Unstapled box and whiskers plot (outliers removed, standard calculation of whiskers)

6 **'Boxplot' - literally**. A three-number summary is given: lower quartile; median; and upper quartile. This plot includes no whiskers and no outliers, thus reducing the disclosure risk particularly relating to extreme observations.

- Safest to report, but losing some insight into skewness that may be of interest.

Fig. 6. Literal box plot (whiskers and outliers removed)

6 Conclusion

This paper explores the disclosure risk when reporting univariate summary statistics.

It has been demonstrated that reporting the mean and standard deviation to summarise a sample can result in a disclosure risk. The risk generally decreases with increasing sample size and as the range of possible values on the measurement scale increases. The R package `uwedragon` can be used to check if reporting the mean and standard deviation for a given sample uniquely identifies the sample values.

To reduce the risk of reconstruction from a sample that uniquely identifies the sample values, noise can be added to summary statistics [5, 7]. In the case of quartiles, the different ways in which these can be calculated, frequently adds what can be described as naturally occurring noise, if the calculation method is not reported.

If concerned about the risk of reporting mean and standard deviation, when $n > 6$ a three-figure summary can instead be reported: median; lower quartile and upper quartile. Although limited to only three values, this can be graphically displayed by the literal 'boxplot' when a basic visualisation of the distribution is desired.

If the sample space is large, and standard deviation is not zero, then the reporting of the mean and standard deviation has a low risk of being fully disclosive of all sample values. However, some indication of extreme values may be apparent for a large standard deviation.

Appendix

See Tables A1, A2, and A3

Table A1. Number of unique solutions, data on 5-point scale.

n	Sample space	Mean and SD reported			Median, Q1 and Q3 reported			Median, Min, Max
		Full	2dp	1dp	SPSS	R	SAS	reported
3	35	33	33	33	35	35	35	35
4	70	56	56	56	70	70	68	41
5	129	87	79	79	48	5	5	5
6	210	105	101	101	151	15	4	9
7	330	131	121	121	0	16	0	5
8	495	141	141	133	39	39	16	9
9	715	177	161	135	0	0	0	5
10	1001	205	181	157	7	39	0	9
11	1365	223	201	130	0	0	0	5
12	1820	243	221	149	7	7	0	9

Table A2. Number of unique solutions, data on 9-point scale.

n	Sample space	Mean and SD reported			Median, Q1 and Q3 reported			Median, Min, Max
		Full	2dp	1dp	SPSS	R	SAS	reported
3	165	145	145	145	165	165	165	165
4	495	271	271	271	493	493	356	145
5	1,287	396	327	286	112	9	9	9
6	3,003	440	364	279	850	44	4	17
7	6,435	527	399	306	0	32	0	9
8	12,870	449	449	284	64	64	16	17
9	24,310	693	499	270	0	0	0	9
10	43,758	701	549	275	0	64	0	17
11	75,582	821	599	246	0	0	0	9
12	125,970	837	649	261	0	0	0	17

Table A3. Number of unique solutions, data on 11-point scale.

n	Sample space	Mean and SD reported			Median, Q1 and Q3 reported			Median, Min, Max
		Full	2dp	1dp	SPSS	R	SAS	reported
3	286	238	238	238	286	286	286	286
4	1,001	443	443	419	971	971	596	221
5	3,003	592	496	386	144	11	11	11
6	8,008	654	530	369	1296	60	4	21
7	19,448	830	580	342	0	40	0	11
8	43,758	652	652	355	64	64	16	21
9	92,378	1080	722	342	0	0	0	11
10	184,756	1044	794	363	0	64	0	21
11	352,716	1263	866	304	0	0	0	11
12	646,646	1232	938	311	0	0	0	21

References

1. Skinner, C.: Statistical disclosure control for survey data. In Handbook of Statistics, vol. 29, pp. 381–396. Elsevier (2009). https://doi.org/10.1016/S0169-7161(08)00015-1
2. Derrick, B., White, P.: Comparing two samples from an individual Likert question. Int. J. Math. Statist. **18**(3) (2017)
3. Derrick, B.: uwedragon: Data Research, Access, Governance Network: Statistical Disclosure Control. R package (2022). https://cran.r-project.org/web/packages/uwedragon/index.html
4. Lowthian, P., Ritchie, F.: Ensuring the confidentiality of statistical outputs from the ADRN. ADRN Technical paper (2017). https://uwe-repository.worktribe.com/output/888435
5. Dinur, I., Nissim, K.: Revealing information while preserving privacy. PODS **2003**, 202–210 (2003)
6. Hozo, S.P., Djulbegovic, B., Hozo, I.: Estimating the mean and variance from the median, range, and the size of a sample. BMC Med. Res. Methodol. **5**(1), 1–10 (2005)
7. Derrick, B., Green, L., Kember, K., Ritchie, F., White, P.: Safety in numbers: Minimum thresholding, Maximum bounds, and Little White Lies: The case of the Mean and Standard Deviation Scottish Economic Society Conference 2022 (2022). www.ses2022.org/sessions/protecting-confidentiality-social-science-research-outputs
8. R Core team: A Language and Environment for Statistical Computing (2021). https://www.R-project.org/
9. Hyndman, R.J., Fan, Y.: Sample quantiles in statistical packages. Am. Stat. **50**(4), 361–365 (1996)
10. Tukey, J.W.: Exploratory Data Analysis, p. 9780201076165. Addison-Wesley, ISBN (1977)